



[Web](#) [Images](#) [Video](#) [News](#) [Maps](#) [more »](#)

duplicate OR replicated representative document

2003

Search

[Advanced Scholar Search](#)

[Scholar Preferences](#)

[Scholar Help](#)

Scholar All articles - [Recent articles](#) Results 1 - 10 of about 3,550 for [duplicate OR replicated representative document OR "web page" OR page webcrawler OR crawler OR spider](#). (0.21 seconds)

Did you mean: [duplicate OR replicated representative document OR "web page" OR page web crawler OR crawler OR spider](#)

[Method and system for detecting duplicate documents in web crawls](#) - all 2 versions »

D Mayerzon, S Shoroff, FG Terek, S Nofin - US Patent 6,547,828, 2003 - [freepatentsonline.com](#)

... **Representative Image**: Method and system for detecting duplicate ... optimize replication by not copying duplicate data. ... An HTML document contains text and metadata ...

[Cited by 4](#) - [Related Articles](#) - [Cachect](#) - [Web Search](#)

[Engineering a multi-purpose test collection for Web retrieval experiments](#) - all 11 versions »

P Bailey, N Craswell, D Hawking - Information Processing and Management, 2003 - Elsevier

... that CRC64 will falsely signal a duplicate in this ... queries for which at least one document from the ... A representative distribution of server sizes was a very ...

[Cited by 163](#) - [Related Articles](#) - [Web Search](#)

[Marie-4: A High-Recall, Self-Improving Web Crawler That Finds Images Using Captions](#) - all 19 versions »

NC Rowe - 2002 - [doi.ieeeecomputersociety.org](#)

... We also eliminate duplicate captions, and only ... candidates and picking three representative keywords from ... the caption-likelihood and document-frequency factors. ...

[Cited by 13](#) - [Related Articles](#) - [Web Search](#) - [Sci Direct](#)

[On the evolution of clusters of near-duplicate Web pages](#) - all 16 versions »

D Fetterly, M Manasse, M Najork - Web Congress, 2003. Proceedings First Latin American, 2003 - [ieeexplore.ieee.org](#)

... of shingles to a small, yet representative, subset. ... each cluster covers all versions of a replicated page. ... found that clusters of near-duplicate documents are ...

[Cited by 46](#) - [Related Articles](#) - [Web Search](#)

[Results from a Web Impact Factor crawler](#) - all 3 versions »

M Treilwall - Journal of Documentation, 2001 - [emeraldinsight.com](#)

... common for servers to allow a document to be ... the pages crawled, indicating that the duplicate pages should ... were chosen because they are representative of the ...

[Cited by 47](#) - [Related Articles](#) - [Web Search](#) - [Sci Direct](#)

[poPi Mirror, mirror on the Web: A study of host pairs with replicated content](#) - all 5 versions »

K Bharata - A Broder - COMPUT. NETWORKS, 1999 - [cambrowski.com](#)

... Host Pairs with Replicated Content ... that almost a third of the Web consists of duplicate pages ... case the samples in the collection may not be very representative. ...

[Cited by 61](#) - [Related Articles](#) - [View as HTML](#) - [Web Search](#)

[Finding replicated Web collections](#) - all 25 versions »

J Cho, N Shivakumar, H Garcia-Molina - ACM SIGMOD Record, 2000 - [portal.acm.org](#)

... of document collections, when the document collections are ... web search engines, by clustering together replicated pages and ... has a node v i for each web page p i ...

[Cited by 51](#) - [Related Articles](#) - [Web Search](#) - [Sci Direct](#)

[Information retrieval on the web](#) - all 23 versions »

M Kobayashi, K Takada - ACM Computing Surveys (CSUR), 2000 - [portal.acm.org](#)

